


## RESEARCH ARTICLE

## Open Access



# An ensemble model of QSAR tools for regulatory risk assessment

Prachi Pradeep<sup>1\*</sup> , Richard J. Povinelli<sup>2</sup>, Shannon White<sup>4</sup> and Stephen J. Merrill<sup>3</sup>

## Abstract

Quantitative structure activity relationships (QSARs) are theoretical models that relate a quantitative measure of chemical structure to a physical property or a biological effect. QSAR predictions can be used for chemical risk assessment for protection of human and environmental health, which makes them interesting to regulators, especially in the absence of experimental data. For compatibility with regulatory use, QSAR models should be transparent, reproducible and optimized to minimize the number of false negatives. In silico QSAR tools are gaining wide acceptance as a faster alternative to otherwise time-consuming clinical and animal testing methods. However, different QSAR tools often make conflicting predictions for a given chemical and may also vary in their predictive performance across different chemical datasets. In a regulatory context, conflicting predictions raise interpretation, validation and adequacy concerns. To address these concerns, ensemble learning techniques in the machine learning paradigm can be used to integrate predictions from multiple tools. By leveraging various underlying QSAR algorithms and training datasets, the resulting consensus prediction should yield better overall predictive ability. We present a novel ensemble QSAR model using Bayesian classification. The model allows for varying a cut-off parameter that allows for a selection in the desirable trade-off between model sensitivity and specificity. The predictive performance of the ensemble model is compared with four in silico tools (Toxtree, Lazar, OECD Toolbox, and Danish QSAR) to predict carcinogenicity for a dataset of air toxins (332 chemicals) and a subset of the gold carcinogenic potency database (480 chemicals). Leave-one-out cross validation results show that the ensemble model achieves the best trade-off between sensitivity and specificity (accuracy: 83.8 % and 80.4 %, and balanced accuracy: 80.6 % and 80.8 %) and highest inter-rater agreement [ $\kappa$  (k): 0.63 and 0.62] for both the datasets. The ROC curves demonstrate the utility of the cut-off feature in the predictive ability of the ensemble model. This feature provides an additional control to the regulators in grading a chemical based on the severity of the toxic endpoint under study.

**Keywords:** Computational toxicology, In silico QSAR tools, Hybrid QSAR models, Ensemble models, Risk assessment

## Background

Chemical risk assessment associated with chemical exposure is necessary for the protection of human and environmental health. Toxicity or adverse effects are major reasons for failure of a potential pharmaceutical, an industrial chemical or a medical device [1–3]. Regulatory risk assessment is the process that ensures marketing of safe and effective drugs, medical devices and other consumer products. Regulatory decisions are primarily

dependent on the short and long term toxic and clinical effects of chemicals. Conventional methods of risk assessment (in vivo experiments and clinical trials) are performed only after product development, and are expensive and time-consuming. Although in vivo experimental studies are the most accurate method for identifying the toxic effects induced by a xenobiotic, time and cost associated with them for new chemical regulation renders them ineffective for regulatory risk assessment.

In silico approaches to predictive toxicology focus on building quantitative structure activity relationship (QSAR) models that can mimic the results of in vivo studies. In silico methods are appealing because they provide a faster alternative to otherwise time-consuming

\*Correspondence: [prachipradeep.15@gmail.com](mailto:prachipradeep.15@gmail.com)

<sup>1</sup> National Center for Computational Toxicology (ORISE Fellow), US EPA, Research Triangle Park, NC, USA

Full list of author information is available at the end of the article

laboratory and clinical testing methods [4, 5]. Currently, several commercial (free or proprietary) and open source in silico QSAR tools are available that can predict the toxic effects of a chemical based on its chemical structure [6, 7]. QSAR models are widely used for identification of chemicals that have a desired biological effect (e.g. drug leads) or for early prediction of potential toxic effects in the pharmaceutical industry. In contrast to industrial use, regulatory use of QSAR models is very different. In a regulatory application, QSAR models can be used to: (1) supplement experimental data, (2) support prioritization in the absence of experimental data, and (3) replace experimental animal testing methods [8, 9].

Several QSAR models have been used and validated by United States (US) regulatory agencies and are rapidly gaining impetus in the European Union (EU) too [10–13]. In the EU, the REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) initiative mandates risk assessment of new and existing chemicals [14]. Similar to REACH, the Organization for Economic Co-operation and Development (OECD), also has a set of internationally agreed upon validation principles for regulatory acceptance of QSAR models [15].

In view of the possible uses of QSAR tools, regulators often use predictions from multiple QSAR tool for arriving at a decision. However, different QSAR tools often make conflicting predictions for a given chemical and also vary in their predictive ability for different classes of chemicals. Often, the validation of a particular QSAR tool and sufficient confidence that it can be used reliably for a given chemical is not available, which makes handling conflicting predictions and determining the best prediction difficult [16]. Transparency in predictions is crucial in developing safety assessment decisions and reports, which makes the use of QSAR tools challenging for a regulatory risk assessment. In this manuscript, we present a Bayesian ensemble model of QSAR tools with improved prediction accuracy and reliability. In the following sections, we discuss the present state-of-the-art, describe ensemble methodology, Bayes classification and present a comparative analysis of the Bayes ensemble model.

## Related work

There are studies that investigate methods for combining predictions from multiple QSAR tools to gain better predictive performance for various toxic endpoints: (1) Several QSAR models were developed and compared using different clustering algorithms (multiple linear regression, radial basis function neural network and support vector machines) to develop hybrid models for bioconcentration factor (BCF) prediction [17]; (2) QSAR models implementing cut-off rules were used to determine

a reliable and conservative consensus prediction from two models implemented in VEGA [18] for BCF prediction [19]; (3) Predictive performance of four QSAR tools (Derek [20, 21], Leadscope [22], MultiCASE [23] and Toxtree [24]) were evaluated and compared to the standard Ames assay [25] for mutagenicity prediction. Pairwise hybrid models were then developed using AND (accepting positive results when both tools predict a positive) and OR combinations (accepting positive results when either one of the tool predicts a positive) [25–27]; (4) A similar AND/OR approach was implemented for the validation and construction of a hybrid QSAR model using MultiCASE and MDL-QSAR [28] tools for carcinogenicity prediction in rodents [29]. The work was extended using more tools (BioEpisteme [30], Leadscope PDM, and Derek) to construct hybrid models using majority consensus predictions in addition to AND/OR combinations [31].

The results of these studies demonstrate that: (1) None of the QSAR tools perform significantly better than others, and they also differ in their predictive performance based upon the toxic endpoint and the chemical datasets under investigation, (2) Hybrid models have an improved overall predictive performance in comparison to individual QSAR tools, and (3) Consensus-positive predictions from more than one QSAR tool improved the identification of true positives. The underlying idea is that each QSAR model brings a different perspective of the complexity of the modeled biological system and combining them can improve the classification accuracy. However, consensus-positive methods are prone to introducing a conservative nature in discarding a potentially non-toxic chemical based on false positive prediction. Therefore, we propose an ensemble learning approach for combining predictions from multiple QSAR tools that addresses the drawbacks of consensus-positive predictions [32, 33]. Hybrid QSAR models using ensemble approaches have been developed for various biological endpoints like cancer classification and prediction of ADMET properties [34–36] but not for toxic endpoints. In this study, a Bayesian ensemble approach is investigated for carcinogenicity prediction, which is discussed in more details in the next section.

## Methods

### QSAR tools

Four open-source QSAR tools were used to make predictions about carcinogenicity for chemicals used in this study:

1. *OECD ToolBox* Chemicals were screened for two mutagenic alerts: (1) in vitro mutagenicity alerts by ISS (Ames mutagenicity), and (2) in vivo mutagenic-

ity alerts by ISS (Micronucleus assay), and two carcinogenic alerts: (1) carcinogenic (genotoxic and non-genotoxic) alerts by ISS, and (2) oncology primary classifications) profiling alerts. A positive result in a profiling category for any chemical substance was considered a positive carcinogenicity prediction for the test chemical [37].

2. *Danish QSAR* Chemicals were screened in the database for mutagenicity, mutagenicity in vivo, and carcinogenicity. A positive or equivalent prediction in any category was recorded as a positive carcinogenicity prediction for the test chemical [38].
3. *Lazar* Chemicals were queried in the tool using the DSSTox carcinogenic potency DBS multicellcall endpoint and the two available mutagenic endpoints (DSSTox carcinogenic potency DBS mutagenicity and Kazius-Bursi Salmonella mutagenicity). A positive result for either category was recorded as a positive carcinogenicity prediction for the test chemical.
4. *Toxtree* Chemicals were queried in the Toxtree using the Benigni/Bossa Rulebase (for mutagenicity and carcinogenicity). If a potential carcinogenic alert based on any QSAR model or if any structural alert for genotoxic and non-genotoxic carcinogenicity was reported, then the prediction was recorded as a positive carcinogenicity prediction for the test chemical.

## Datasets

Two datasets, that consist of both carcinogenic and non-carcinogenic chemicals, were used for training and testing:

1. *Air toxins* A set of chemicals potentially emitted in the industrial environment was obtained from the Western Australia Department of Health. The dataset consists of 332 chemicals with a carcinogen to non-carcinogen ratio of 114:218.
2. *Gold carcinogenic potency database (CPDB)* The CPDB houses results from chronic, long-term animal cancer tests on a variety of chemicals [39]. The database was screened for all chemicals with positive or negative carcinogenic data in both male and female mice and/or rats. A chemical was considered carcinogenic if the response in either species and/or gender had a TD50 data, else it was considered non-carcinogenic. The final dataset consists of 480 chemicals with a carcinogen to non-carcinogen ratio of 258:222.

Selection of chemicals in each dataset was based on the availability of experimental in vivo carcinogenicity data [obtained from the Carcinogenic Potency Database and

Chemical Carcinogenesis Research Information System (CCRIS [40])] and predictions from all four QSAR tools. The list of chemicals in both the datasets is provided in Tables 1 and 2 of the Additional file 1.

## Bayes ensemble model

A Bayes ensemble model is based on the concept of prior probabilities [41, 42]. The model uses training data for classification by estimating uncertain quantities using the Bayes theorem. Bayes theorem uses the training data as evidence (E) for a seen outcome (O), to construct a probability for predicting the outcome when the evidence is seen in the future [43]. The probability of observing the outcome in the past (training dataset) is termed as the prior probability ( $P(O)$ ) and the probability of predicting the outcome occurring in the future is termed as the posterior probability ( $P(O|E)$ ). The Bayes theorem calculates the posterior probability using Eq. (1).

$$P(O|E) = \frac{P(E|O)P(O)}{P(E)}, \quad (1)$$

where  $P(O)$  is the probability of the outcome and  $P(E)$  is the probability of the evidence. In a binary classification problem, the final predicted class is the one with a higher value of  $P(O|E)$ .

In this study, the training data consisted of predictions from four QSAR tools and true experimental class about the nature of the chemical (carcinogenic or non-carcinogenic). Each tool was used to make a prediction about the class ( $\omega$ ), which is recorded as 1 or 0 representing carcinogenic and non-carcinogenic, respectively. Since there were four QSAR tools, the possible number of combinations of predictions is  $k = 2^4 (= 16)$ . Each unique prediction combination is represented by the vector  $s_k$ . The posterior probability of a chemical being carcinogenic ( $\omega = 1$ ) or non-carcinogenic ( $\omega = 0$ ) associated with each prediction combination,  $P(\omega|s = s_k)$  is then calculated using Eq. (2).

$$P(\omega|s = s_k)_k = \frac{P(s_k|\omega)P(\omega)}{P(s_k)}, \quad (2)$$

where,  $s_k$  is the prediction combination for the test chemical,  $P(s_k|\omega)$  is the prior probability of observing a prediction combination  $s_k$  given that a chemical is carcinogenic or non-carcinogenic,  $P(\omega)$  is the probability of a chemical being carcinogenic or non-carcinogenic and  $P(s_k)$  is the probability of a particular prediction combination from the QSAR tools. So, for each prediction combination ( $s_k$ ) there is an associated posterior probability ( $P(\omega|s = s_k)$ ), which is used to make the final classification ( $\omega$ ). The algorithm was implemented in Matlab R2015a [44] and the source code is provided in the Additional file 2.

### Algorithm

The approach outlined above is implemented in two steps for estimation of the final classification ( $\omega'$ ):

Step 1. The posterior probability of a test chemical being carcinogenic was calculated from Eq. (3) and was used to construct a decision table as shown in Table 1 for both datasets.

$$P(\omega = 1|s = s_k) = \frac{P(s_k|\omega = 1)P(\omega = 1)}{P(s_k)} \quad (3)$$

where,

$$P(s_k|\omega = 1) = \frac{N_{(\omega=1,s_k)}}{N_{(\omega=1)}}, \quad (4)$$

$$P(\omega = 1) = \frac{N_{(\omega=1)}}{N}, \text{ and} \quad (5)$$

$$P(s_k) = \frac{N_{s_k}}{N}. \quad (6)$$

So,

$$P(\omega = 1|s = s_k) = \frac{\left(\frac{N_{(\omega=1,s_k)}}{N_{(\omega=1)}}\right)\left(\frac{N_{(\omega=1)}}{N}\right)}{\left(\frac{N_{s_k}}{N}\right)} \quad (7)$$

$$= \frac{N_{(\omega=1,s_k)}}{N_{s_k}}, \quad (8)$$

**Table 1 Prediction combination table with posterior probability,  $P(\omega|s = s_k)$ , for each combination number,  $s_k$ , which represents a prediction combination from each of the four QSAR tools**

Combination number	Tool 1	Tool 2	Tool 3	Tool 4	Posterior probability
$s_1$	0	0	0	0	$P(\omega s = s_1)$
$s_2$	0	0	0	1	$P(\omega s = s_2)$
$s_3$	0	0	1	0	$P(\omega s = s_3)$
$s_4$	0	0	1	1	$P(\omega s = s_4)$
$s_5$	0	1	0	0	$P(\omega s = s_5)$
$s_6$	0	1	0	1	$P(\omega s = s_6)$
$s_7$	0	1	1	0	$P(\omega s = s_7)$
$s_8$	0	1	0	1	$P(\omega s = s_8)$
$s_9$	0	1	1	1	$P(\omega s = s_9)$
$s_{10}$	1	0	0	0	$P(\omega s = s_{10})$
$s_{11}$	1	0	0	1	$P(\omega s = s_{11})$
$s_{12}$	1	0	1	0	$P(\omega s = s_{12})$
$s_{13}$	1	0	1	1	$P(\omega s = s_{13})$
$s_{14}$	1	1	0	0	$P(\omega s = s_{14})$
$s_{15}$	1	1	0	1	$P(\omega s = s_{15})$
$s_{16}$	1	1	1	1	$P(\omega s = s_{16})$

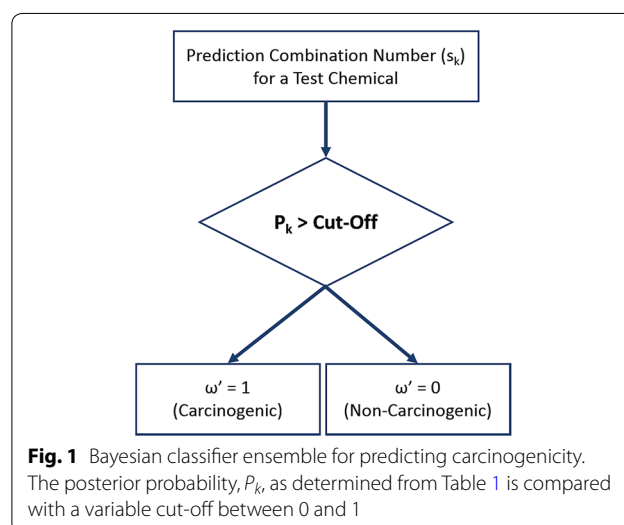
where  $N_{s_k}$  was the number of chemicals with a prediction combination  $s_k$  in the training dataset,  $N_{(\omega=1)}$  was the total number of carcinogens in the training dataset,  $N_{(\omega=1,s_k)}$  was the number of carcinogens with prediction combination  $s_k$ ,  $N$  was the total number of chemicals in the training dataset and  $k$  ranges from 1 to 16. Tables 3 and 4 in the Additional file 1 list the number of samples in each of the 16 prediction classes for both datasets.

Step 2. For a new test chemical, the prediction combination vector  $s_k$  was determined and was used to look up the posterior probability  $P(\omega = 1|s = s_k)$  or  $P_k$  associated with it from the decision table. The final prediction ( $\omega'$ ) was estimated based on the value of  $P_k$ , which was compared to a variable *cut-off* as outlined in Fig. 1. The cut-off represents the value of posterior probability beyond which a new test chemical can be classified as carcinogenic. The value of the cut-off can be varied (between 0 and 1) leading to different decision points for the final classification.

The Bayes ensemble model is, thus, very powerful in giving a user the flexibility of adjusting the cut-off to reach a desired level of sensitivity and specificity as demonstrated in the results. The flexibility in changing the cut-off also makes the model endpoint independent.

### Model validation

One of the major concerns with the use of QSAR tools for a regulatory purpose is the reliability in their predictions. QSAR tools need to be assessed for their scientific validity so that regulatory organizations have a sound scientific basis for decision making. The OECD member



countries agreed upon a set of principles as guidelines for scientifically validating a QSAR model. In accordance with these guidelines, external model validation was performed and a range of model statistics were calculated for a comprehensive performance analysis. The leave one out cross validation (LOOCV) technique was used for external validation where  $N$  models were developed each with  $(N - 1)$  chemicals as training set and 1 chemical as the test set. The following standard metrics were then calculated to assess the performance of the models:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}, \quad (9)$$

$$\text{Sensitivity (SN)} = \frac{TP}{TP + FN}, \quad (10)$$

$$\text{Specificity (SP)} = \frac{TN}{TN + FP}, \quad (11)$$

$$\text{Balanced Accuracy (BA)} = \frac{SP + SN}{2} \quad (12)$$

$$\text{PPV} = \frac{TP}{TP + FP}, \text{ and} \quad (13)$$

$$\text{NPV} = \frac{TN}{TN + FN}, \quad (14)$$

where TP is the number of true positives, TN is the number true negatives, FP is the number of false positives, and FN is the number of false negatives reported in the tests. Accuracy or concordance is a measure of correctness of overall predictions. Sensitivity is a measure of correctness in prediction of positives or toxic chemicals and specificity is a measure of correctness in prediction of negatives or non-toxic chemicals. Balanced accuracy (BA) is the arithmetic mean of sensitivity and specificity and represents a trade-off between the two values. Positive predictive value (PPV) is the proportion of positives or toxic chemicals that are correctly predicted and negative predictive value (NPV) is the proportion of negatives or non-toxic chemicals that are correctly predicted. High sensitivity or low false negatives is especially important under REACH requirements. BA, PPV and NPV are crucial in understanding the predictive power of the models based on the representation of carcinogenic and non-carcinogenic chemicals in the training datasets.

The OECD guidelines emphasize appropriate measures of goodness-of-fit, robustness and predictivity of QSAR models. Several reports discuss potential techniques for internal and external measures of model validation [45–47]. Therefore, in addition to the standard metrics two

conceptually simpler statistical parameters are also calculated, which are indicative of overall concordance and performance of each model as compared to chance and each other:

1. *Cohen's Kappa ( $\kappa$ )* The Kappa coefficient is a measure of pairwise inter-rater agreement or specific agreement compared to a chance agreement. It is calculated as below:

$$\kappa = \frac{(TP + TN) - \left( \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{N} \right)}{1 - \left( \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{N} \right)}. \quad (15)$$

In this study, the Kappa coefficient is used to compare how well the predictions from various tools agree with the experimental or true values. Values of  $\kappa = 0$ ,  $0.41 < \kappa < 0.60$ ,  $0.61 < \kappa < 0.80$  and  $\kappa = 1$  represent no, moderate, substantial and perfect agreement, respectively [48, 49].

2. *Receiver Operating Characteristics (ROC) Curve* A ROC curve is a plot of true positive rate (sensitivity) and the false positive rate (1-specificity). A ROC curve demonstrates how the performance of a binary classifier changes as the threshold parameters are varied [50]. Area under the ROC curve can be used to compare the classification tools; higher area implies a better classification.

## Results and discussion

### Accuracy, sensitivity, specificity, balanced accuracy, PPV and NPV

Statistical performance of the ensemble model in comparison to the various QSAR tools is summarized in Tables 2 and 3. The statistics for the Bayes ensemble model are presented for three different cut-offs, which demonstrate the utility of the cut-off feature. As shown, the accuracy (>80 %), balanced accuracy (>78 %), PPV (>79 %) and NPV (>79 %) of the Bayes ensemble model is highly improved compared to the base classifiers (QSAR tools) for both the datasets. The specificity was substantially improved which adheres with the REACH legislations emphasis on the reduction of false negatives.

The statistics demonstrate the inability of any particular QSAR tool to make consistent predictions across different chemical datasets. In case of Bayes ensemble model, varying the cut-off leads to perfect sensitivity (cut-off = 0) or perfect specificity (cut-off = 1). However, cut-off values of 0.4, 0.5 and 0.6 result in a balanced sensitivity and specificity with only a minor change in all calculated statistics for both the datasets. This demonstrates the robustness of the Bayes ensemble model. Additionally,



**Table 2 Performance metrics for air toxins dataset**

Model	Accuracy (%)	SN (%)	SP (%)	BA (%)	PPV (%)	NPV (%)	Kappa ( $\kappa$ )
Toxtree	75.56	68.18	79.51	73.85	64.10	82.32	0.47
Lazar	75.24	74.55	75.61	75.08	62.12	84.70	0.48
Danish QSAR	74.29	80.91	70.73	75.82	59.73	87.35	0.48
OECD toolbox	76.19	69.09	80.00	74.55	64.96	82.83	0.48
Bayes ensemble (Cut-off = 0.4)	83.81	70.00	91.22	80.61	81.05	85.00	0.63
Bayes ensemble (Cut-off = 0.5)	83.81	70.00	91.22	80.61	81.05	85.00	0.63
Bayes ensemble (Cut-off = 0.6)	82.22	65.45	91.22	78.34	80.00	83.11	0.59

**Table 3 Performance metrics for the CPDB dataset**

Model	Accuracy (%)	SN (%)	SP (%)	BA (%)	PPV (%)	NPV (%)	Kappa ( $\kappa$ )
Toxtree	66.04	84.50	44.59	64.55	63.93	71.22	0.30
Lazar	80.63	86.05	74.32	80.19	79.57	82.09	0.61
Danish QSAR	65.00	91.09	34.68	62.89	61.84	77.00	0.27
OECD toolbox	64.79	84.50	41.89	63.20	62.82	69.93	0.27
Bayes ensemble (Cut-off = 0.4)	81.04	83.33	75.23	79.28	80.14	82.27	0.62
Bayes ensemble (Cut-off = 0.5)	80.21	84.50	75.23	79.87	79.85	80.68	0.60
Bayes ensemble (Cut-off = 0.6)	80.42	84.50	77.03	80.77	80.83	79.91	0.61

pairwise Student's t-tests was used to establish statistically significant differences between the predictions from the Bayes ensemble model (cut-off = 0.5) and all the tools at a 5 % significance level, for both datasets.

#### Cohen's Kappa coefficient

As seen in Tables 2 and 3, the Bayes ensemble model has the highest Kappa coefficient. This means that the Bayes ensemble predictions best concur with the experimental data. Toxtree, Danish QSAR and the OECD Toolbox demonstrate less than moderate agreement with the experimental values for both the datasets. The Bayes ensemble model with cut-off = 0.4 has a  $\kappa > 0.62$  for both the datasets. It is an indication of stronger and more substantial agreement with the experimental values compared to the other QSAR tools.

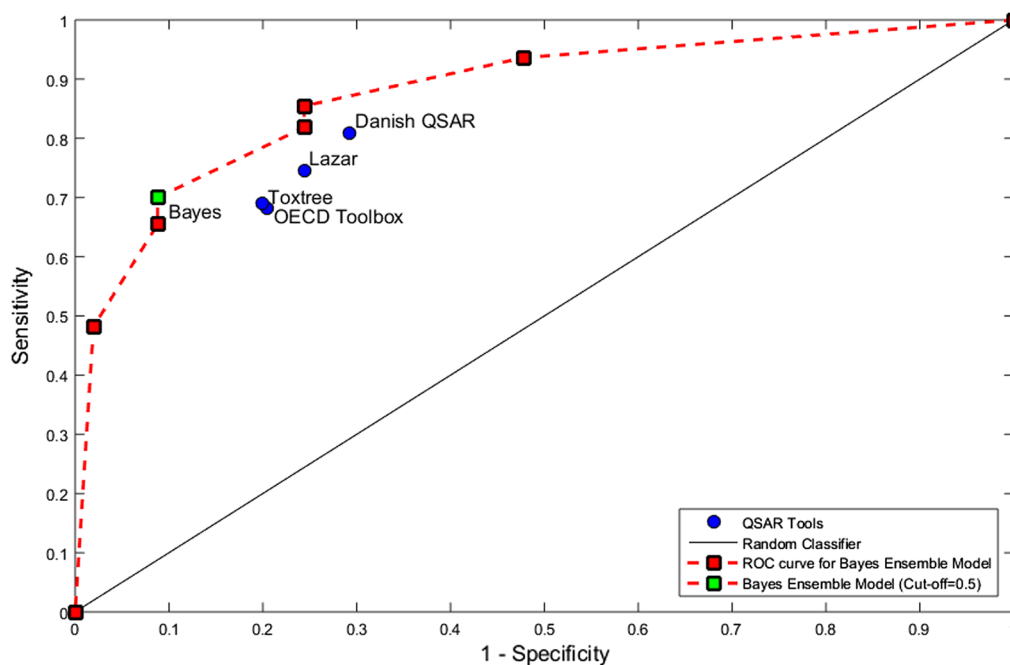
#### ROC curve

Figure 2 shows the receiver operating characteristics plot for all the QSAR tools and the Bayes ensemble model. An ideal binary predictor would have zero false predictions and so the desired point on the ROC curve is top left corner where sensitivity is one and (1-specificity) is zero. The black line corresponds to

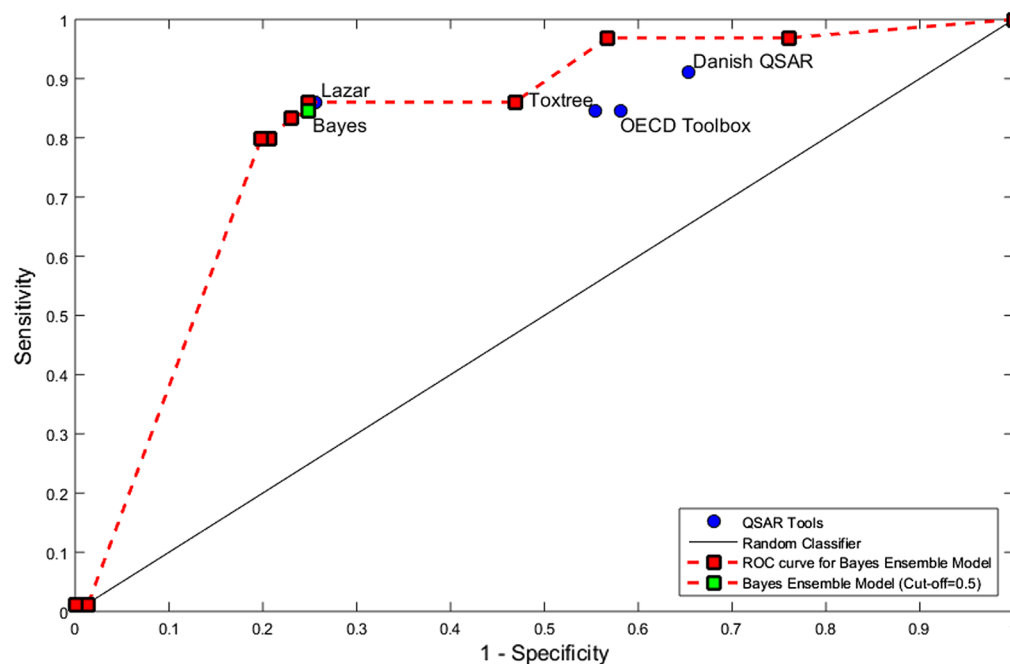
the performance of a random classifier, which does not have any preferences in binary outcomes. The higher the area under the ROC curve, the greater is the predictive ability of the model. The tools give a binary prediction, therefore, they are represented as a point on the ROC plot. In the case of Bayes ensemble model, a curve can be traced for each sensitivity-specificity combination obtained after changing the value of the cut-off. In this study, the cut-off is varied between 0 and 1 with a step size of 0.1 allowing for 11 decision points for model validation. Hence, the ROC plot consists of data points corresponding to each value of cut-off, which can be traced to obtain a ROC curve. The ROC curve for the Bayes ensemble model is higher than all the other tools implying better quality of predictions.

*The variable cut-off in the ROC curve can be adjusted to select a trade-off between sensitivity and specificity. This feature provides an additional control to the regulating agencies in grading a chemical based on the severity of the toxic endpoint under study.* It exhibits user-control and flexibility in the predictive ability of the ensemble model.

Overall, the results show that the Bayes ensemble model is better and more consistent with respect to



### **a** Air Toxins Dataset



### **b** CPDB Dataset

**Fig. 2** Receiver operator characteristics (ROC) curve of Bayes ensemble model as compared to other QSAR tools. The Bayes model at different thresholds is depicted by *red points*, at 0.5 cut-off by *green point* and the base QSAR tools by *blue points*. The ROC plot for the Bayes ensemble model is depicted by the *red dotted line*. **a** Air toxins dataset, **b** CPDB dataset

different *in silico* QSAR tools. The model combines predictions from various *in silico* tools in a transparent and reproducible manner. It can also be optimized to reduce

the number of false predictions while maintaining flexibility in addressing other considerations in making these predictions.

## Conclusion

The results of this study demonstrate that different QSAR tools vary in the quality of predictions depending on the underlying algorithm and training datasets. Ensemble machine learning presents a new approach for combining the predictions from multiple QSAR tools. The strength of an ensemble model depends on the diversity of the algorithm and the predictive ability of the base models. Each individual tool has its strengths and weaknesses and an ensemble model enables leveraging the benefits of individual tools, minimizing the impact of their algorithmic differences and increase in chemical space coverage.

The Bayes ensemble model presented here is consistent in its performance across both the datasets. The results specifically show improved (1) accuracy and balanced accuracy in the predictions, (2) specificity and positive predictive value, which are an indication of reduction in false positive predictions, and (3) Kappa coefficient, across both the datasets. The statistics demonstrate how ensemble machine learning methods can be used to increase the capability of consensus QSAR models for toxicity prediction.

The Bayes ensemble model offers flexibility in achieving a desired trade-off between sensitivity and specificity. It also demonstrates how multiple QSAR tools with different complexity and accuracy can be used together for developing more reliable predictors. The results suggest that ensemble modeling techniques are a good strategy for refining hybrid models and to tailor their use based on the severity and concerns associated with the toxic endpoint under study. We presented an example application with Toxtree, Lazar, OECD Toolbox, and Danish QSAR, and two different classes of chemical datasets for carcinogenicity prediction. This approach can be extended to different tools and different kinds and sizes of chemical datasets for different toxic endpoints as well.

## Additional files

**Additional file 1: Section S1.** Dataset 1: Air Toxins. **Section S2.** Dataset 2: Subset of Carcinogenic Potency Database. **Section S3.** Distribution of prediction combinations.

**Additional file 2.** Matlab generated code file and is self-explanatory.

## Authors' contributions

PP conceived and designed the study, implemented the analysis, and drafted the manuscript. RJP and SJM advised with the study design. SW helped in compiling the datasets. All authors read and approved the final manuscript.

## Author details

<sup>1</sup> National Center for Computational Toxicology (ORISE Fellow), US EPA, Research Triangle Park, NC, USA. <sup>2</sup> Electrical and Computer Engineering Department, Marquette University, Milwaukee, WI, USA. <sup>3</sup> Department of Mathematics, Statistics, and Computer Science, Marquette University, Milwaukee, WI, USA. <sup>4</sup> Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA.

## Acknowledgements

This project was supported in part by an appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration. Sincere thanks to Ronald Brown (Toxicologist, CDRH) for his help in identification and development of the datasets used in this study and for his valuable suggestions in the development of this work.

## Competing interests

The authors declare that they have no competing interests.

Received: 21 April 2016 Accepted: 7 September 2016

Published online: 22 September 2016

## References

1. Kramer JA, Sagartz JE, Morris DL (2007) The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov* 6:636–649
2. Coecke S, Pelkonen O, Leite SB, Bernauer U, Bessems JG, Bois FY, Gundert-Remy U, Loizou G, Testai E, Zaldivar JM (2013) Toxicokinetics as a key to the integrated toxicity risk assessment based primarily on non-animal approaches. *Toxicol In-vitro* 27:1570–1577
3. Kaplan AV, Baim DS, Smith JJ, Feigal DA, Simons M, Jefferys D, Thomas JF, Kuntz RE, Leon MB (2004) Medical device development from prototype to regulatory approval. *Circulation* 109(25):3068–3072
4. Muster W, Breidenbach A, Fischer H, Kirchner S, Miller L, Phler A (2008) Computational toxicology in drug development. *Drug Discov Today* 13(7):303–310
5. Dearden JC (2003) In silico prediction of drug toxicity. *J Comput Aided Mol Des* 17(2–4):119–127
6. Serafimova R, Gatnik MF, Worth A (2010) Review of qsar models and software tools for predicting genotoxicity and carcinogenicity. Publications Office of the European Union. JRC Scientific and technical reports
7. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) Review of methods for qsar applicability domain estimation by the training set. Technical report. The European Commission—Joint Research Centre Institute for Health & Consumer Protection-ECVAM
8. Cronin MT, Jaworska JS, Walker JD, Comber MH, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ Health Perspect* 111(10):1391–1401
9. Hfer T, Gerner I, Gundert-Remy U, Liebsch M, Schulte A, Spielmann H, Vogel R, Wettig K (2004) Animal testing and alternative approaches for the human health risk assessment under the proposed new european chemicals regulation. *Arch Toxicol* 78(10):549–564
10. Valerio LG Jr (2011) In silico toxicology models and databases as FDA Critical Path Initiative toolkits. *Hum Genomics* 5(3):200–207
11. Zeeman M, Auer CM, Clements RG, Nabholz JV, Boethling RS (1995) US EPA regulatory perspectives on the use of QSAR for new and existing chemical evaluations. *SAR QSAR Environ Res* 3(3):179–201
12. US EPA (2014) Estimation programs interface suite for microsoft windows. United States Environmental Protection Agency, Washington
13. EPA T.E.S.T. <http://www.epa.gov/nrmrl/std/qsar/qsar.html>. Last accessed Apr 2014
14. Worth AP (2010) Recent advances in QSAR studies. Springer Netherlands, pp 367–382. Chap. 13: "The role of QSAR methodology in the regulatory assessment of chemicals"
15. Jaworska JS, Comber M, Auer C, Van Leeuwen CJ (2003) Summary of a workshop on regulatory acceptance of (q) sars for human health and environmental endpoints. *Environ Health Perspect* 111(10):1358–1360
16. Gleeson MP, Modi S, Bender A, Marchese L, Robinson R, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC (2012) The challenges involved in modeling toxicity data in silico: a review. *Curr Pharm Des* 18(9):1266–1291
17. Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73(11):1701–1707



18. VEGA. <http://www.vega-qsar.eu/>. Last accessed Apr 2013
19. Gissi A, Nicolotti O, Carotti A, Gadaleta D, Lombardo A, Benfenat E (2013) Integration of qsar models for bioconcentration suitable for reach. *Sci Total Environ* 456:325–332
20. Sanderson DM, Earnshaw CG (1991) Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum Exp Toxicol* 10(4):261–273
21. Lhasa DEREK. <http://www.lhasalimited.org/products/derek-nexus.htm>. Last accessed Oct 2014
22. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE (2000) Leadscope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 40(6):1302–1314
23. MultiCASE Inc. <http://multicase.com/>
24. Toxtree—Toxic Hazard Estimation by decision tree approach. <http://tox-tree.sourceforge.net/>. Last accessed 2013
25. Ames BN, McCann J, Yamasaki E (1975) Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutat Res* 31:347–364
26. Benignia R, Giulianib A (1988) Computer-assisted analysis of interlaboratory ames test variability. *J Toxicol Environ Health* 25:135–148
27. Hillebrecht A, Muster W, Brigo A, Kansy M, Weiser T, Singer T (2011) Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. *Chem Res Toxicol* 24(6):843–854
28. MDL QSAR, MDL information systems. <http://www.mdli.com>
29. Contrera JF, Kruhlak NL, Matthews EJ, Benz RD (2007) Comparison of MC4PC and MDL-QSAR rodent carcinogenicity predictions and the enhancement of predictive performance by combining QSAR models. *Regul Toxicol Pharmacol* 49(3):172–182
30. Valencia A (2010) BioEpisteme—an in silico approach for predicting and understanding the underlying molecular mechanisms contributing to toxicity responses. *Toxicol Lett* 196(S25):1–48. doi:10.1016/j.toxlet.2010.03.117
31. Matthews EJ, Kruhlak NL, Benz RD, Contrera JF, Marchant CA, Yang C (2008) Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol Mech Methods* 18(2–3):189–206
32. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 51:181–207
33. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J Chem Inf Model* 45:786–799
34. Tan AC, Gilbert D (2003) Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinform* 2:75–83
35. Oza NC, Tumer K (2008) Classifier ensembles: select real-world applications. *Inf Fusion* 9:4–20
36. Pissurlenkar RR, Khedkar VM, Iyer RP, Coutinho EC (2011) Ensemble QSAR: a QSAR method based on conformational ensembles and metric descriptors. *J Comput Chem* 32(10):2204–2218
37. OECD QSAR Toolbox. <http://www.qsartoolbox.org/>
38. Danish (Q)SAR Database. <http://qsar.food.dtu.dk/>
39. Gold LS (1984) The Carcinogenic Potency Project (CPDB). <http://toxnet.nlm.nih.gov/cpdb/>. Last accessed Oct 2014
40. Chemical Carcinogenesis Research Information System (2014) <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>. Last accessed Mar 2014
41. Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. The MIT Press, Cambridge
42. Kuncheva LI, Rodriguez JJ (2014) A weighted voting framework for classifiers ensembles. *Knowl Inf Syst* 38:259–275
43. Valpola H, Honkela A, Karhunen J, Raiko T, Giannakopoulos X, Ilin A, Oja (2001) Bayesian ensemble learning of generative models. Biennial Report of Adaptive Informatics Research Center, Aalto Univ
44. MATLAB R2012a (2012) The MathWorks Inc
45. Chirico N, Gramatica P: Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J Chem Inf Model* 51(9):2320–2335
46. Keefer CE, Kauffman GW, Gupta RR (2013) Interpretable, probability-based confidence metric for continuous quantitative structureactivity relationship models. *J Chem Inf Model* 53(2):368–383
47. Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predictive ability by external validation techniques. *J Chemom* 24(3–4):194–201
48. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
49. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. *Fam Med* 37(5):360–363
50. Zou KH, OMalley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 115(5):654–657

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)